

ΚΕΦΑΛΑΙΟ ΔΕΥΤΕΡΟ

Εξερευνώντας τα δεδομένα μας-Περιγραφική Στατιστική

Το πρώτο βήμα στην ανάλυση ενός συνόλου δεδομένων, που αποτελούν μετρήσεις ενός δείγματος είναι η παρουσίαση και σύνοψη των πληροφοριών του δείγματος για τις μεταβλητές που περιλαμβάνονται σε αυτό, χρησιμοποιώντας μεθόδους της Περιγραφικής Στατιστικής. Το S.P.S.S. έχει ενσωματωμένες διαδικασίες για το σκοπό αυτό τόσο για ποιοτικές όσο και για ποσοτικές μεταβλητές. Στις ενότητες που ακολουθούν παραθέτουμε τη διαδικασία για τη συνοπτική παρουσίαση ποιοτικών δεδομένων και έπειτα ποσοτικών δεδομένων.

2.1 Ποιοτικές μεταβλητές

Η συνοπτική παρουσίαση των δεδομένων μίας ποιοτικής μεταβλητής (βλέπε σχετικά Ζωγράφος, 2003, σελ. 18-31, 41-43) επιτυγχάνεται α) με τον πίνακα συχνοτήτων των δεδομένων και β) με τις γραφικές τους παραστάσεις (ραβδόγραμμα, κυκλικό διάγραμμα).

Ο πίνακας συχνοτήτων μιας ποιοτικής μεταβλητής προκύπτει από την απαρίθμηση και καταγραφή των δειγματικών τιμών στην αντίστοιχη κατηγορία. Ένας ολοκληρωμένος πίνακας συχνοτήτων μίας ποιοτικής μεταβλητής περιλαμβάνει τη στήλη των Συχνοτήτων (η συχνότητα παριστάνει τον αριθμό των φορών που μία κατηγορία της ποιοτικής μεταβλητής εμφανίζεται στο δείγμα) και τη στήλη των Σχετικών συχνοτήτων (η σχετική συχνότητα παριστάνει το ποσοστό επί τοις εκατό των φορών εμφάνισης μίας τιμής στο δείγμα). Επιπλέον, μπορούν να συμπεριληφθούν στον πίνακα συχνοτήτων διατάξιμων μόνο ποιοτικών μεταβλητών, η στήλη των Αθροιστικών συχνοτήτων (παριστάνει το πλήθος των τιμών του δείγματος που είναι μικρότερες ή το πολύ ίσες από μία τιμή) και η στήλη των Αθροιστικών σχετικών συχνοτήτων (παριστάνει το ποσοστό επί τοις εκατό των τιμών του δείγματος που είναι μικρότερες ή ίσες από μία τιμή).

Ένας τρόπος άμεσης κατανόησης των χαρακτηριστικών της κατανομής των συχνοτήτων επιτυγχάνεται με μία ειδική γραφική παράσταση που ονομάζεται ραβδόγραμμα. Στον οριζόντιο άξονα ενός ραβδογράμματος συχνοτήτων (εναλλακτικά ενός ραβδογράμματος σχετικών συχνοτήτων) σημειώνονται οι κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται, ενώ στον κατακόρυφο άξονα οι αντίστοιχες συχνοτήτες (εναλλακτικά οι αντίστοιχες σχετικές συχνοτήτες).

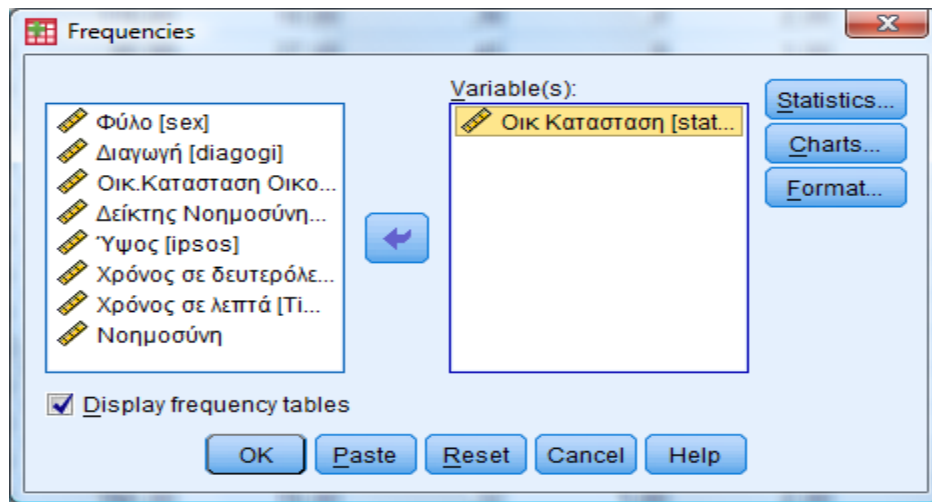
Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος χωρισμένος σε τομείς, όσες και οι κατηγορίες στις οποίες τα μέλη του πληθυσμού κατατάσσονται. Το εμβαδό κάθε τομέα απεικονίζει το ποσοστό των ατόμων που ανήκουν στην αντίστοιχη κατηγορία.

Υλοποίηση στο S.P.S.S.

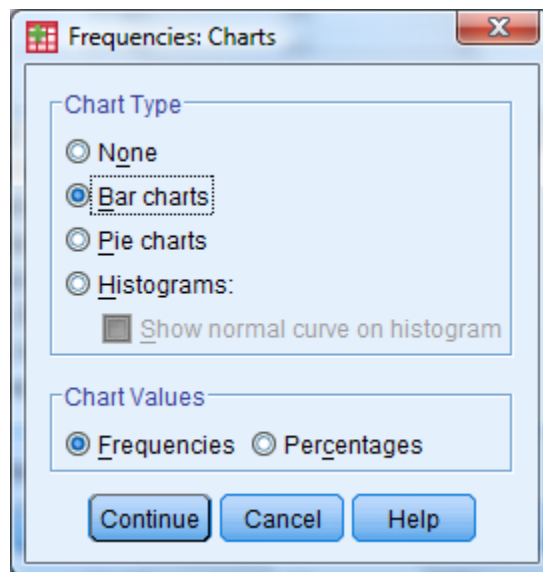
Σε συνέχεια του Παραδείγματος 1.1 θα γίνει ο πίνακας συχνοτήτων, το ραβδόγραμμα και το κυκλικό διάγραμμα της μεταβλητής που περιγράφει την οικονομική κατάσταση της οικογένειας.

Η συνοπτική παρουσίαση των δεδομένων ποιοτικών μεταβλητών γίνεται με την ακόλουθη διαδικασία:

1. Analyze→Descriptive Statistics→Frequencies.
2. Στο νέο παράθυρο διαλόγου που προκύπτει επιλέγουμε τις προς ανάλυση ποιοτικές μεταβλητές και τις μεταφέρουμε στο κουτί Variable(s). Έχοντας επιλέξει μόνο το πλαίσιο Display frequency tables θα παραχθούν μόνο οι πίνακες συχνοτήτων.



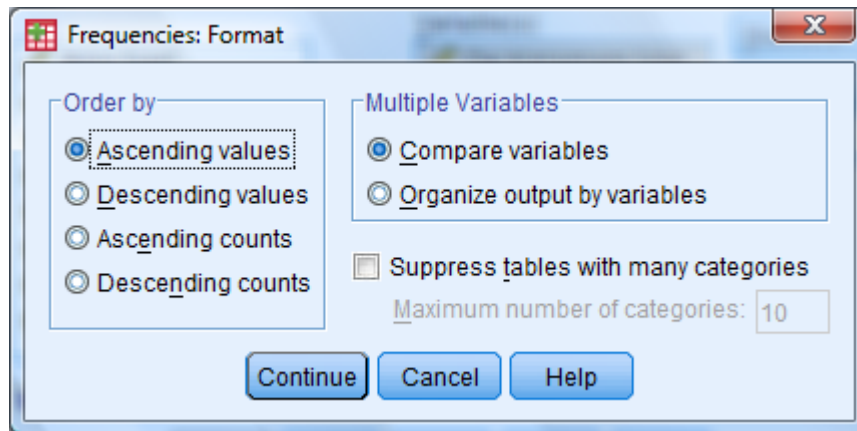
3. Από την επιλογή Charts μπορούμε να κατασκευάσουμε: Ραβδογράμματα (Bar charts), Κυκλικά Διαγράμματα (Pie charts). Τα ιστογράμματα (Histograms), όπως θα δούμε και στην επόμενη ενότητα, αφορούν την περίπτωση ποσοτικών μεταβλητών. Δυστυχώς κάθε φορά έχουμε τη δυνατότητα μίας επιλογής μεταξύ του Bar Charts και Pie Charts. Επιλέγοντας π.χ. την κατασκευή ραβδογράμματος (ή κυκλικού διαγράμματος), ενεργοποιείται η επιλογή Chart Values από όπου επιλέγοντας Frequencies ή Percentages καθορίζουμε αν στον κατακόρυφο άξονα των υπό κατασκευή ραβδογραμμάτων ή κυκλικών διαγραμμάτων θα εμφανίζονται οι απόλυτες συχνότητες (Frequencies) ή οι σχετικές συχνότητες (Percentages), αντίστοιχα.



4. Τέλος, από την επιλογή Format του κεντρικού παραθύρου διαλόγου Frequencies καθορίζουμε αν ο πίνακας συχνοτήτων θα εμφανιστεί είτε σε αύξουσα ή φθίνουσα σειρά εμφάνισης των διαφορετικών κατηγοριών της ποιοτικής μεταβλητής (Order by Ascending or Descending values) είτε σύμφωνα με τη συχνότητα εμφάνισης των διαφορετικών κατηγοριών (Order by Ascending or Descending Counts).

Επιπλέον, αν στο πλαίσιο Variable(s) του κεντρικού παραθύρου διαλόγου Frequencies έχουν δηλωθεί περισσότερες από μία μεταβλητές μπορούμε είτε να αποκτούμε τα αποτελέσματα σε ένα πίνακα για όλες (Compare Variables) είτε να γίνεται η ανάλυση ξεχωριστά για καθεμία (Organize output by variables).

Τέλος, η επιλογή Suppress tables with more than n categories εμποδίζει την εμφάνιση πινάκων με περισσότερες από n κατηγορίες (που είναι ο μέγιστος αριθμός κατηγοριών που δηλώνεται στο πλαίσιο Maximum number of categories).



Ερμηνεία αποτελεσμάτων

Statistics

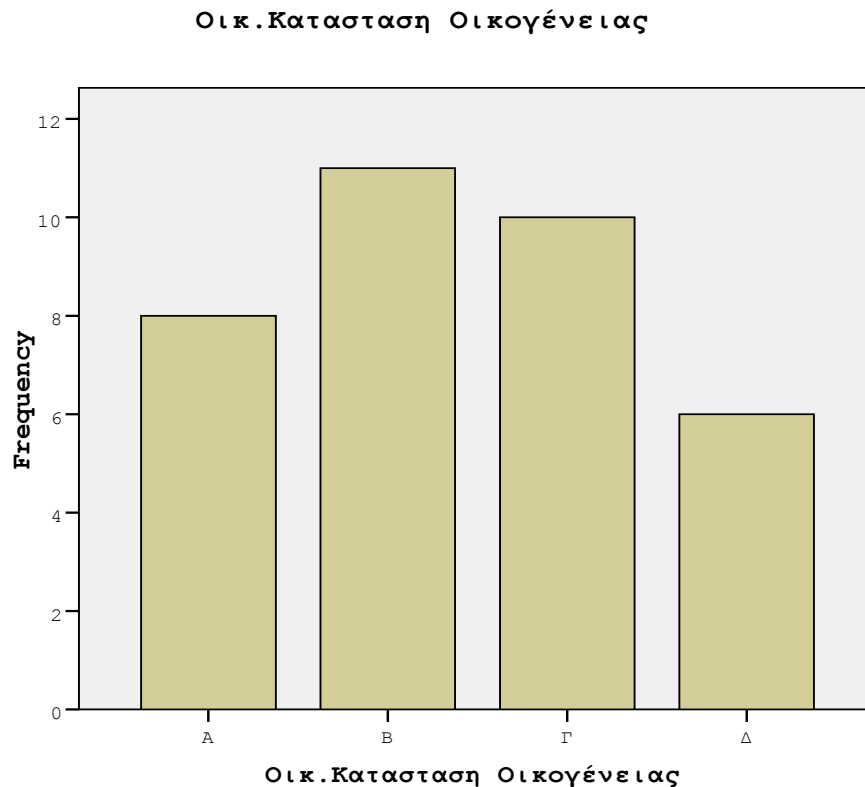
Οικ.Κατασταση Οικογένειας		
N	Valid	35
	Missing	0

Οικ.Κατασταση Οικογένειας

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	8	22,9	22,9	22,9
	B	11	31,4	31,4	54,3
	Γ	10	28,6	28,6	82,9
	Δ	6	17,1	17,1	100,0
	Total	35	100,0	100,0	

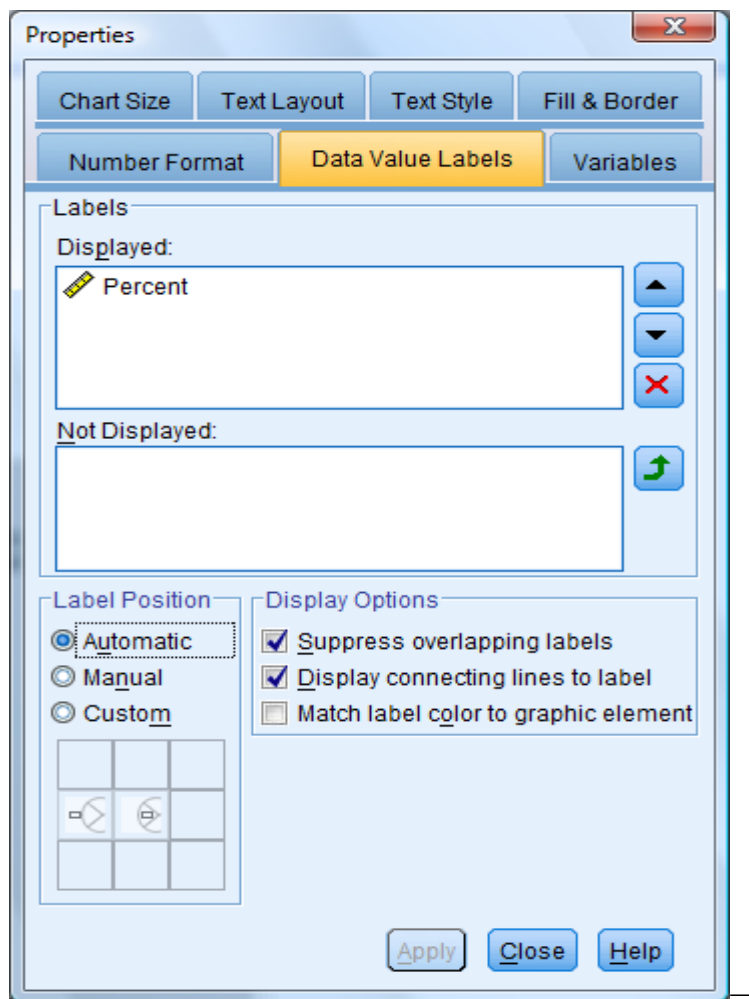
Από τον πρώτο πίνακα πληροφορούμαστε ότι οι διαθέσιμες δειγματικές τιμές είναι 35 (Valid=35) και δεν υπάρχουν ελλιπείς τιμές (Missing=0). Ο δεύτερος πίνακας είναι ουσιαστικά ο πίνακας συχνοτήτων για την Οικονομική Κατάσταση της οικογένειας. Έτσι, από τη στήλη Frequency (Στήλη Συχνοτήτων) προκύπτει ότι 8, 11, 10 και 6 από τις 35 συνολικά οικογένειες είναι οικονομικής κατάστασης Α, Β, Γ, Δ αντίστοιχα. Επιπλέον, από τη στήλη Percent (Στήλη Σχετικών Συχνοτήτων) έχουμε π.χ. ότι 22,9% των ερωτηθέντων ανήκει στην κατηγορία Α. Επισημαίνεται ότι το ποσοστό στη στήλη Percent υπολογίζεται στο σύνολο των ερωτηθέντων συμπεριλαμβανομένου και των πιθανών ελλιπών τιμών. Από την άλλη μεριά το ποσοστό στη στήλη Valid Percent υπολογίζεται στο σύνολο αυτών που έχουν απαντήσει. Εδώ προφανώς προκύπτει ισότητα

καθώς δεν έχουμε ελλειπείς παρατηρήσεις. Τέλος, από τη στήλη Cumulative Percent (Στήλη Αθροιστικών Σχετικών Συχνοτήτων) προκύπτει για παράδειγμα ότι 82,9% των ερωτηθέντων έχουν εισόδημα μικρότερο ή ίσο των 900 Ευρώ (Γ=600-900 Ευρώ). Η στήλη αυτή όπως γνωρίζουμε από τη θεωρία έχει νόημα μόνο για διατάξιμες ποιοτικές μεταβλητές, όπως αυτή του παραδείγματος. Τέλος, έχουμε το παρακάτω ραβδόγραμμα.



Κάνοντας διπλό κλικ στο ραβδόγραμμα ή στο κυκλικό διάγραμμα που προκύπτει στο Output (δηλαδή στο παράθυρο των αποτελεσμάτων) έχουμε τη δυνατότητα περαιτέρω επεξεργασίας του (ως προς το χρώμα, τον τρόπο εμφάνισης, τους τίτλους, τους υπότιτλους κ.ά.).

Ειδικότερα, θέλοντας να εμφανίζονται τα ποσοστά της κάθε κατηγορίας στο κυκλικό διάγραμμα κάνουμε διπλό κλικ σε αυτό και στο νέο παράθυρο επιλέγουμε Elements→Show Data Labels και στο επόμενο παράθυρο διαλόγου κάτω από το πλαίσιο Displayed ζητούμε να εμφανίζεται το Percent.



2.2 Ποσοτικές μεταβλητές

Η συνοπτική παρουσίαση των δεδομένων ποσοτικών μεταβλητών περιλαμβάνει τον υπολογισμό των τιμών διάφορων στατιστικών μέτρων, όπως η μέση τιμή (Mean), η τυπική απόκλιση (Std Deviation), οι συντελεστές κύρτωσης και λοξότητα (Kurtosis, Skewness, αντίστοιχα), η διάμεσος (median), η επικρατούσα τιμή (mode), το εύρος (range), τα ποσοστιαία σημεία (Percentile values) κ.ά. Το δεύτερο στάδιο περιλαμβάνει την πιθανή κατασκευή του ιστογράμματος (histogram) και θηκογράμματος (boxplot) της υπό εξέταση ποσοτικής μεταβλητής, τον έλεγχο της ύπαρξης ακραίων τιμών στις δειγματικές τιμές της υπό εξέταση μεταβλητής, καθώς και τον έλεγχο αν οι διαθέσιμες δειγματικές τιμές μπορούν να θεωρηθούν ότι προέρχονται από έναν πληθυσμό που

περιγράφεται ικανοποιητικά από την κανονική κατανομή (βλέπε σχετικά Ζωγράφος, 2003, σελ. 45-55).

Περιγραφικά μέτρα

Μια συνοπτική παρουσίαση των δεδομένων ποσοτικών μεταβλητών επιτυγχάνεται με τα περιγραφικά μέτρα, που διακρίνονται σε μέτρα θέσης και μέτρα διασποράς.

Ένα μέτρο θέσης είναι μία αριθμητική τιμή ενδεικτική της θέσης, του σημείου γύρω από το οποίο ένα σύνολο δεδομένων συγκεντρώνεται. Τέτοια είναι η μέση τιμή \bar{X} (μέσος όρος των μετρήσεων), η διάμεσος (η τιμή εκείνη που χωρίζει τα δεδομένα σε δύο ίσα μέρη έτσι ώστε το πλήθος των μετρήσεων που βρίσκονται αριστερά της να είναι ίσο με το πλήθος των μετρήσεων που βρίσκεται δεξιά της) και η επικρατούσα τιμή ή κορυφή (η τιμή με τη μεγαλύτερη συχνότητα).

Ένα μέτρο διασποράς είναι μία αριθμητική τιμή ενδεικτική του τρόπου με τον οποίο τα δεδομένα κατανέμονται γύρω από τη μέση τιμή. Τέτοια μέτρα είναι το εύρος (παριστάνει τη διαφορά της ελάχιστης από τη μέγιστη τιμή), η διακύμανση S^2 (εκφράζει τη μεταβλητότητα ενός συνόλου αριθμητικών δεδομένων από τη μέση τους τιμή), η τυπική απόκλιση S (η θετική τετραγωνική ρίζα της διακύμανσης).

Άλλα περιγραφικά μέτρα, μεταξύ άλλων, είναι ο συντελεστής λοξότητας και κύρτωσης αντίστοιχα, που μετρούν την ασυμμετρία της κατανομής και την

«αιχμηρότητα» της, αντίστοιχα. Ορίζονται από τις σχέσεις $b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{nS^3}$ και

$b_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{nS^4}$, αντίστοιχα, όπου X_1, \dots, X_n είναι οι διαθέσιμες δειγματικές τιμές (n

το μέγεθος του δείγματος), $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ η δειγματική μέση τιμή και

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ η δειγματική διακύμανση.}$$

Με βάση την λοξότητα οι κατανομές διακρίνονται σε συμμετρικές όταν $b_1 = 0$ (σε αυτές ανήκει η κανονική κατανομή), σε θετικά ασύμμετρες (ή λοξές δεξιά) όταν $b_1 > 0$ και σε αρνητικά ασύμμετρες (ή λοξές αριστερά) όταν $b_1 < 0$.

Με βάση την κύρτωση οι κατανομές διακρίνονται σε λεπτόκυρτες όταν $b_2 > 3$, σε μεσόκυρτες όταν $b_2 = 3$ (σε αυτές ανήκει η κανονική κατανομή) και σε πλατύκυρτες όταν $b_2 < 3$. Το λογισμικό του S.P.S.S. υπολογίζει την τιμή $b_2 - 3$, έτσι ώστε η σύγκριση και η εξέταση για ενδείξεις αποκλίσεων από την κανονικότητα να γίνεται με το μηδέν.

γ) Τα εκατοστιαία ποσοστιαία σημεία. Το p-οστό εκατοστιαίο σημείο έχει την ιδιότητα p% των μετρήσεων να είναι μικρότερες ή ίσες από αυτό, και τέλος,

δ) τα τεταρτημόρια που έχουν την ιδιότητα να χωρίζουν το σύνολο των μετρήσεων σε τέσσερα ίσα μέρη και δεν είναι τίποτε άλλο από το 25°, 50°, 75° ποσοστιαίο σημείο.

Ιστόγραμμα συχνοτήτων

Πολλές φορές οι τιμές μιας ποσοτικής μεταβλητής είναι πολυάριθμες και για τη συνοπτική παρουσίασή τους κρίνεται σκόπιμη η ομαδοποίησή τους. Οι ομάδες έχουν τη μορφή κλειστών συνεχόμενων διαστημάτων. Το ιστόγραμμα συχνοτήτων συνίσταται από ένα σύνολο συγγενών ορθογώνιων παραλληλόγραμμων, των οποίων το ύψος είναι ανάλογο με τη συχνότητα κάθε ομάδας και το μήκος τους ανάλογο με το μήκος της ομάδας. Οι τιμές της μεταβλητής (ουσιαστικά τα άκρα των ομάδων) τοποθετούνται στον οριζόντιο άξονα, ενώ οι συχνότητες στον κατακόρυφο άξονα.

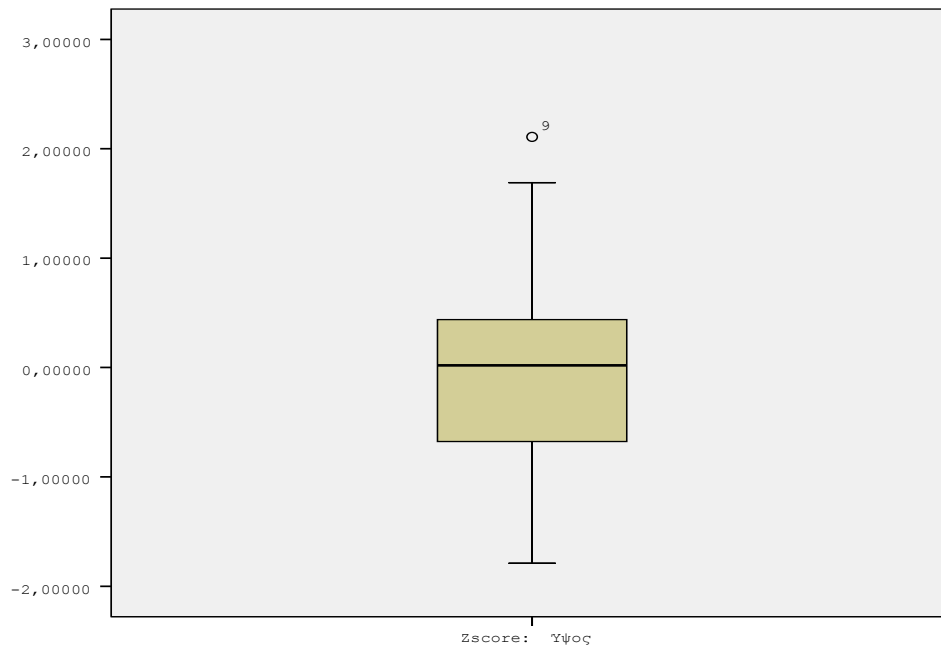
Φυλλογράφημα

Μία παραλλαγή του ιστογράμματος είναι το φυλλογράφημα (stem and leaf plot). Το φυλλογράφημα δεν είναι τίποτε άλλο παρά το αποτέλεσμα της περιστροφής κατά 90°

του ιστογράμματος συχνοτήτων. Επομένως γίνεται αντιληπτό ότι το φυλλογράφημα επιδεικνύεται προς μία μεριά (δεξιά). Το μήκος κάθε γραμμής αντιστοιχεί στον αριθμό των παρατηρήσεων που ανήκουν στο διάστημα. Η κύρια διαφοροποίηση του φυλλογραφήματος σε σχέση με το ιστόγραμμα συχνοτήτων είναι ότι αναπαρίσταται κάθε τιμή με μία αληθινή τιμή.

Θηκόγραμμα

Στο θηκόγραμμα παριστάνονται περιγραφικά μέτρα όπως η διάμεσος, το 25^ο και 75^ο ποσοστιαίο σημείο και οι ακραίες τιμές («αντιφατικές» τιμές σε σχέση με τις υπόλοιπες παρατηρούμενες τιμές του συνόλου δεδομένων).



Το κάτω άκρο του κουτιού είναι το 25^ο ποσοστιαίο σημείο και το πάνω άκρο το 75^ο. Η διάμεσος παριστάνεται από μία οριζόντια γραμμή μέσα στο κέντρο του κουτιού. Στην αρχή και στην κορυφή του σχήματος σημειώνονται δύο οριζόντιες γραμμές, που αναφέρονται ως φράχτες (whiskers). Το θηκόγραμμα μας βοηθά στο να δούμε αν υπάρχουν ακραίες τιμές (τιμές πέρα από τους whiskers, επισημαίνονται με «ο» και είναι ακραίες, ενώ με * επισημαίνονται οι extreme) καθώς και πιθανές αποκλίσεις από την κανονική κατανομή (αν η διάμεσος είναι πιο κοντά στην κορυφή ή στην αρχή του κουτιού και όχι στο κέντρο). Ο άνω και κάτω φράκτης καθορίζονται από τις σχέσεις

1^ο τεταρτημόριο – 1.5 * ενδοτεταρτημοριακό εύρος

και

3^ο τεταρτημόριο +1.5 * ενδοτεταρτημοριακό εύρος

αντίστοιχα, όπου το ενδοτεταρτημοριακό εύρος είναι η διαφορά του 3^{ου} από το 1^ο τεταρτημόριο.

Παρατήρηση: Ο όρος ακραία τιμή αναφέρεται σε μία παρατήρηση η οποία κατά μία έννοια είναι «αντιφατική» σε σχέση με τις υπόλοιπες παρατηρούμενες τιμές του συνόλου δεδομένων. Οι ακραίες τιμές αρχικά θα πρέπει να επισημαίνονται και αφού διαπιστωθεί ότι δεν πρόκειται για λάθη κατά την πληκτρολόγηση των δεδομένων να μελετώνται. Δεν συνιστάται ο αυτόματος αποκλεισμός τους από την έρευνα χωρίς καμία διάκριση, καθώς πολλές φορές και οι ακραίες τιμές περικλείουν εξίσου σημαντικές πληροφορίες. Επισημαίνεται ότι κάθε φορά αποφασίζουμε για την ύπαρξη μίας ακραίας τιμής και αφού την αποκλείσουμε προβαίνουμε σε ύπαρξης επιπρόσθετης ακραίας τιμής. Η μεθοδολογία αυτή θα αναπτυχθεί διεξοδικά σε επόμενη ενότητα καθώς και στο Κεφάλαιο 4.

Έλεγχος κανονικότητας

Η υπόθεση της κανονικότητας είναι μία από τις υποθέσεις πάνω στις οποίες έχει θεμελιωθεί η στατιστική συμπερασματολογία. Οι περισσότερες από τις μεθοδολογίες της Παραμετρικής Στατιστικής υποθέτουν, προϋποθέτουν ότι τα δεδομένα προέρχονται από έναν πληθυσμό, ο οποίος περιγράφεται ικανοποιητικά από την κανονική κατανομή. Για το λόγο αυτό πολλοί τρόποι ελέγχου έχουν εμφανιστεί στη βιβλιογραφία για την υπόθεση της κανονικότητας, τόσο στατιστικοί όσο και γραφικοί. Από τους στατιστικούς τρόπους ελέγχου ξεχωρίζει το στατιστικό τεστ που προτάθηκε από τους Shapiro-Wilk και οι επεκτάσεις αυτού. Η βασική γραφική μέθοδος για τον έλεγχο της κανονικότητας είναι το Q-Q (quantile-quantile) γράφημα, το οποίο συγκρίνει τα ποσοστιαία σημεία (quantile) του δείγματος έναντι των πληθυσμιακών ποσοστιαίων σημείων της κανονικής κατανομής. Αν τα σημεία είναι κοντά σε ευθεία γραμμή δεν υπάρχει ένδειξη για απόκλιση από την κανονικότητα. Παρεκκλίσεις από την ευθεία γραμμή δηλώνουν μη κανονικότητα. Ο τύπος της μη γραμμικότητας μπορεί να υποδηλώνει το τρόπο

απόκλισης από την κανονικότητα. Το S.P.S.S. για κάθε Q-Q γράφημα που κατασκευάζει μας δίνει και μία γραφική παράσταση που ονομάζεται Detrended Q-Q Plot. Η γραφική αυτή μέθοδος δείχνει τις ατομικές αποκλίσεις μεταξύ παρατηρούμενων και εκτιμώμενων αθροιστικών τιμών (ή εκατοστημορίων). Τα σημεία αυτά κατανέμονται γύρω από μία οριζόντια γραμμή που αντιστοιχεί στο 0.

Παρατήρηση 1: Στα στατιστικά πακέτα η απόφαση για την αποδοχή ή απόρριψη μιας στατιστικής υπόθεσης δεν γίνεται εξετάζοντας αν η τιμή του στατιστικού ανήκει στην **περιοχή απόρριψης** (γνωστή και ως **κρίσιμη περιοχή**), αλλά στη βάση των p-τιμών (p-value ή Sig.) Η **p-τιμή** ενός στατιστικού τεστ είναι η μικρότερη τιμή του επιπέδου σημαντικότητας για την οποία απορρίπτεται η μηδενική υπόθεση. Εύκολα προκύπτει τότε ότι **απορρίπτουμε την προς έλεγχο μηδενική υπόθεση αν η p-τιμή είναι μικρότερη από το προκαθορισμένο επίπεδο σημαντικότητας (συνήθως το 0.05).**

Παρατήρηση 2: Έστω Y_1, \dots, Y_n είναι οι n το πλήθος διαθέσιμες δειγματικές τιμές της υπό μελέτη μεταβλητής, οι οποίες αποκλίνουν από την κανονικότητα. Ο μετασχηματισμός Box-Cox (βλέπε Box and Cox (1964)) δίνεται από τη σχέση

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda \left(\dot{Y} \right)^{\lambda-1}}, & \lambda \neq 0 \\ \left(\dot{Y} \right) \ln(Y_i), \dots, \lambda = 0 \end{cases}$$

όπου $\dot{Y} = (Y_1 \cdot Y_2 \cdot \dots \cdot Y_n)^{1/n}$ και υποθέτει ότι για κάποια τιμή της παραμέτρου λ τα μετασχηματισμένα δεδομένα ικανοποιούν την υπόθεση της κανονικότητας.

Υλοποίηση στο S.P.S.S.

Για την υλοποίηση των παραπάνω μπορούν να χρησιμοποιηθούν δύο διαδικασίες του λογισμικού, οι διαδικασίες Descriptives και Frequencies, η καθεμία εκ των οποίων μας δίνει διαφορετικές δυνατότητες και επιλογές. Αν έχουμε όμως ως στόχο την πιο αναλυτική παρουσίαση των δεδομένων μας χρησιμοποιούμε μία πιο σύνθετη διαδικασία,

την διαδικασία Explore. Στη συνέχεια θα παραθέσουμε τον τρόπο υλοποίησης των παραπάνω με τη διαδικασία Explore και απλώς θα αναφέρουμε τις επιπλέον δυνατότητες που δίνουν οι άλλες δύο διαδικασίες.

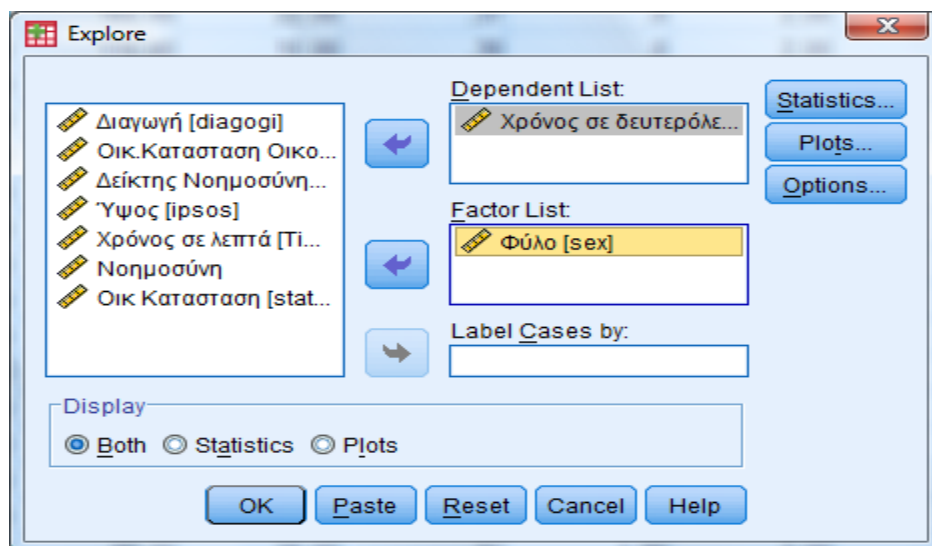
Διαδικασία Explore

Η διαδικασία Explore μπορεί να χρησιμοποιηθεί για την απόκτηση πλήθους στατιστικών μέτρων καθώς και γραφικών παραστάσεων τόσο για το σύνολο των δεδομένων όσο και ξεχωριστά για κατηγορίες αυτών.

Χωρίς βλάβη της γενικότητας στη συνέχεια περιγράφεται η μεθοδολογία που ακολουθείται αν το ενδιαφέρον επικεντρώνεται στη συνοπτική παρουσίαση και αρχική μελέτη του χρόνου σε δευτερόλεπτα που διανύει ένα παιδί τα 100 μέτρα ως προς το φύλο του.

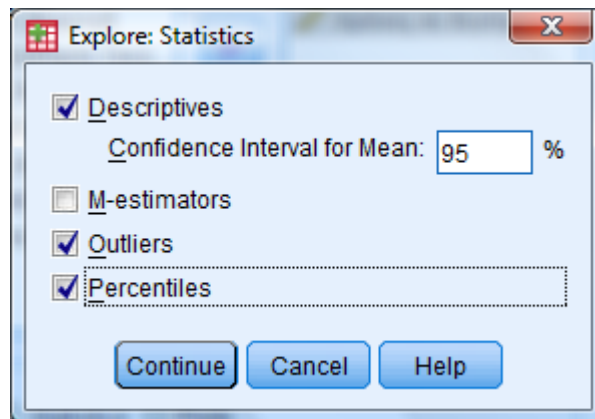
Από το κεντρικό παράθυρο διαλόγου επιλέγουμε:

1. Analyze → Descriptive Statistics → Explore.
2. Στο παράθυρο διαλόγου που προκύπτει τοποθετούμε στο πλαίσιο Dependent τις ποσοτικές (υποχρεωτικά και μόνο) μεταβλητές που θέλουμε να αναλύσουμε. Στο πλαίσιο Factor list τοποθετούμε τις πιθανές ποιοτικές-κατηγορικές μεταβλητές (και μόνο) ως προς τις κατηγορίες των οποίων θέλουμε να προχωρήσουμε την ανάλυση μας, π.χ. έτος σπουδών, φύλο κ.ο.κ.



Επιπρόσθετα διατηρώντας την προεπιλογή Display Both (στο κάτω αριστερό άκρο του παραθύρου) έχουμε τη δυνατότητα απόκτησης τόσο στατιστικών μέτρων όσο και γραφημάτων. Το πλαίσιο Label Cases By το αφήνουμε ως έχει κενό, έτσι ώστε το S.P.S.S να χρησιμοποιήσει την προεπιλογή του αύξοντα αριθμού παρατήρησης.

3. Από την επιλογή Statistics επιλέγουμε τα ακόλουθα



Descriptives (προεπιλογή): απόκτηση των κυριότερων περιγραφικών μέτρων, όπως η διάμεσος, η μέση τιμή, η τυπική απόκλιση κ.ά. καθώς και ενός π.χ. 95% διαστήματος εμπιστοσύνης για την πληθυσμιακή μέση τιμή του υπό μελέτη χαρακτηριστικού (που έχει δηλωθεί στο πλαίσιο Dependent List). Το διάστημα αυτό υπολογίζεται υπό την υπόθεση της κανονικότητας. Επομένως χρειάζεται προσοχή στην περίπτωση αποκλίσεων από την κανονικότητα.

Outliers: το λογισμικό θα μας δώσει τις πέντε μικρότερες και πέντε μεγαλύτερες τιμές κάθε μεταβλητής που έχει δηλωθεί στο πλαίσιο Dependent List, ως προς τις κατηγορίες της μεταβλητής που έχει δηλωθεί στο πλαίσιο Factor List.

Percentiles: υπολογίζει το λογισμικό το 5^ο –95^ο ποσοστιαίο σημείο.

4. Από την επιλογή Plots έχουμε τη δυνατότητα για τα ακόλουθα:

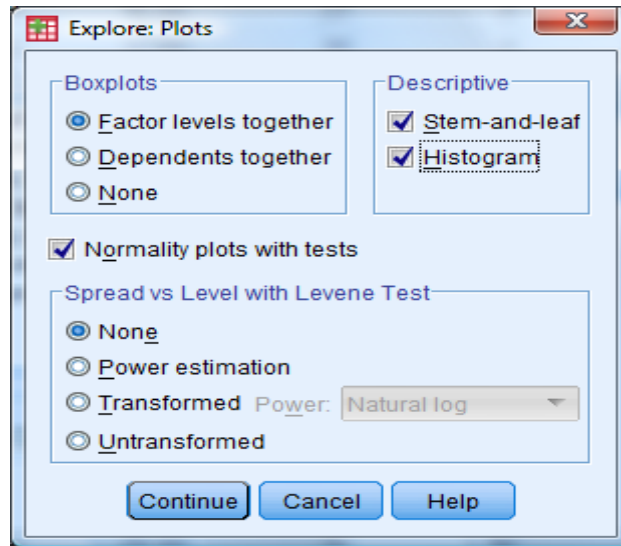
Boxplots: αποκτούμε τα θηκογράμματα. Σημειώνουμε επιπρόσθετα ότι η επιλογή *Factor levels together* δημιουργεί ένα ξεχωριστό πλαίσιο θηκογραμμάτων για καθεμία μεταβλητή που έχουμε δηλώσει στο πλαίσιο Dependent Variables, ως προς κάθε κατηγορία της ποιοτικής μεταβλητής που έχουμε δηλώσει στο πλαίσιο Factor List, ενώ η επιλογή *Dependents together* δημιουργεί ένα ξεχωριστό πλαίσιο θηκογραμμάτων για καθεμία κατηγορία της ποιοτικής μεταβλητής που έχουμε δηλώσει στο πλαίσιο Factor

list ως προς καθεμία από τις ποσοτικές μεταβλητές που έχουν δηλωθεί στο πλαίσιο Dependent Variable. Είναι προτιμότερο να επιλέγουμε το Factor levels together.

Descriptive: έχουμε διαθέσιμες τις επιλογές *Steam-and-Leaf* και *Histogram*, από όπου δηλαδή μπορούμε να αποκτήσουμε το φυλλογράφημα και το ιστόγραμμα για τις ποσοτικές μεταβλητές.

Με την επιλογή Normality plots with tests αποκτούμε τόσο γραφικούς τρόπους ελέγχου της κανονικότητας (normal probability και detrended normal probability plots) όσο και στατιστικά τεστ ελέγχου (το Kolmogorov-Smirnov στατιστικό, με τη διόρθωση του Lilliefors καθώς και το Shapiro-Wilk στατιστικό τεστ, το οποίο και είναι προτιμότερο να εμπιστευόμαστε).

Spread vs. Level with Levene Test: μας δίνει τρόπο να ελέγξουμε την υπόθεση ότι η εξαρτημένη μεταβλητή (έχει δηλωθεί στο πλαίσιο Dependent List) έχει την ίδια διακύμανση μέσα σε δύο ή περισσότερους πληθυσμούς (που προκύπτουν από τις κατηγορίες της μεταβλητής που υπεισέρχονται στο πεδίο Factor List). Η παραπάνω υπόθεση της ισότητας των διακυμάνσεων ή ομοσκεδαστικότητας, όπως θα αναφερθεί σε επόμενα κεφάλαια, είναι αρκετά σημαντική για την εφαρμογή κάποιων μεθοδολογιών. Ο έλεγχος αυτής της υπόθεσης επιτυγχάνεται με το στατιστικό τεστ του *Levene* και επιλέγοντας το *Untransformed data*. Αν η ισότητα απορριφθεί, επαναλαμβάνοντας τα παραπάνω βήματα, επιλέγοντας το πλαίσιο *Power Estimation* το λογισμικό μας προσδιορίζει τον καλύτερο μετασχηματισμό. Έπειτα χρησιμοποιώντας από το πλαίσιο *Transformed* τον μετασχηματισμό που μας έχει υποδειχθεί πρωτύτερα θα πάρουμε το καινούριο γράφημα και θα πραγματοποιηθεί ο έλεγχος της ομοσκεδαστικότητας για τα μετασχηματισμένα δεδομένα.



5. Από την επιλογή Options καθορίζουμε τον τρόπο χειρισμού των ελλιπών τιμών. Για τους τρόπους χειρισμού των ελλιπών δεδομένων έχει γραφεί πληθώρα ερευνητικών εργασιών και συγγραμμάτων. Στα πλαίσια του μαθήματός μας θα αναφέρουμε ότι θα διατηρούμε την (προ)επιλογή *Exclude cases listwise*. Η τεχνική αυτή χειρισμού των ελλιπών δεδομένων περιορίζει την ανάλυση σε εκείνες τις πειραματικές μονάδες (γραμμές) όπου είναι διαθέσιμες οι παρατηρούμενες τιμές σε όλες τις υπό μελέτη μεταβλητές (στήλες).

Ερμηνεία αποτελεσμάτων

Case Processing Summary

Φύλο	Cases						
	Valid		Missing		Total		
	N	Percent	N	Percent	N	Percent	
Χρόνος σε δευτερόλεπτα	Αγόρι	19	100,0%	0	,0%	19	100,0%
	Κορίτσι	16	100,0%	0	,0%	16	100,0%

Ο πίνακας αυτός μας πληροφορεί ότι από τους 35 συμμετέχοντες 19 ήταν αγόρια και 16 κορίτσια χωρίς να υπάρχουν ελλειπείς τιμές.

Στον πίνακα Descriptives μας δίνονται διάφορα περιγραφικά μέτρα (και όχι μόνο) για τη μεταβλητή που περιγράφει το χρόνο σε δευτερόλεπτα που διένυσαν τα 100 μέτρα. Χρρίζουν ιδιαίτερης προσοχής και σχολιασμού τα ακόλουθα.

Descriptives

Φύλο		Statistic	Std. Error	
Χρόνος σε δευτερόλεπτα	Αγόρι	Mean	24,8947	
		95% Confidence Interval for Mean	Lower Bound	22,9017
			Upper Bound	26,8878
		5% Trimmed Mean	24,8830	
		Median	25,0000	
		Variance	17,099	
		Std. Deviation	4,13514	
		Minimum	18,00	
		Maximum	32,00	
		Range	14,00	
		Interquartile Range	6,00	
		Skewness	-,106	,524
		Kurtosis	-,943	1,014
		Κορίτσι	Κορίτσι	Mean
95% Confidence Interval for Mean	Lower Bound			20,4188
	Upper Bound			24,2062
5% Trimmed Mean	22,0694			
Median	21,5000			
Variance	12,629			
Std. Deviation	3,55375			
Minimum	18,00			
Maximum	31,00			
Range	13,00			
Interquartile Range	5,75			
Skewness	,959			,564
Kurtosis	,765			1,091

Η μέση τιμή (Mean) του χρόνου στα αγόρια είναι μεγαλύτερη από ότι στα κορίτσια (24,8947 έναντι 22,315). Το λογισμικό μας δίνει το 95% διάστημα εμπιστοσύνης (95% Confidence Interval for Mean, Lower and Upper Bound) το οποίο είναι αξιόπιστο με την προϋπόθεση ότι δεν υπάρχουν ακραίες τιμές και τα δεδομένα του χρόνου σε δευτερόλεπτα για κάθε ένα από τους δύο πληθυσμούς (αγόρια και κορίτσια) προέρχονται από πληθυσμούς που περιγράφονται από την κανονική κατανομή.

Από τους συντελεστές λοξότητας και κύρτωσης δεν μπορούμε να αποφανθούμε για το αν τα δεδομένα προέρχονται από κανονική κατανομή, καθώς οι τιμές αυτές δεν αποκλίνουν πολύ από το μηδέν. Επομένως απαιτούνται περισσότεροι γραφικοί και κυρίως στατιστικοί τρόποι ελέγχου της υπόθεσης της κανονικότητας.

Παρατηρούμε ότι ο μέσος χρόνος σε δευτερόλεπτα τόσο των αγοριών όσο και των κοριτσιών είναι περίπου ίσος με τη διάμεσο (median) του χρόνου, επομένως τα δεδομένα μπορούν να θεωρηθούν ότι προέρχονται από συμμετρικό πληθυσμό.

Επιπλέον, στον πίνακα Percentiles εμφανίζονται τα ποσοστιαία σημεία, ενώ στη στήλη Extreme Values οι χρόνοι των 5 πιο αργών και πιο γρήγορων στα 100 μέτρα αγοριών και κοριτσιών.

Percentiles

Φύλο			Percentiles						
			5	10	25	50	75	90	95
Weighted Average(Definitio n 1)	Χρόνος σε δευτερόλεπτα	Αγόρι	18	18	22	25	28	30,	.
		Κορίτσι	18	18,7	19	21,5	24,75	28,2	.
Tukey's Hinges	Χρόνος σε δευτερόλεπτα	Αγόρι			22	25,	28		
		Κορίτσι			19	21,5	24,5		

Extreme Values

Φύλο				Case Number	Value
Χρόνος σε δευτερόλεπτα	Αγόρι	Highest	1	28	32,00
			2	11	30,00
			3	34	30,00
			4	21	29,00
			5	20	28,00(a)
	Κορίτσι	Lowest	1	32	18,00
			2	9	18,00
			3	6	20,00
			4	5	21,00
			5	8	22,00(b)
Χρόνος σε δευτερόλεπτα	Αγόρι	Highest	1	25	31,00
			2	16	27,00
			3	2	25,00
			4	33	25,00
			5	13	24,00(c)
	Κορίτσι	Lowest	1	3	18,00
			2	26	19,00
			3	22	19,00
			4	15	19,00
			5	4	19,00

a Only a partial list of cases with the value 28,00 are shown in the table of upper extremes.

b Only a partial list of cases with the value 22,00 are shown in the table of lower extremes.

c Only a partial list of cases with the value 24,00 are shown in the table of upper extremes.

Στον πίνακα Tests of Normality αποφασίζουμε για την αποδοχή ή όχι της υπόθεσης της κανονικότητας με βάση τις p-τιμές του ελέγχου που δίνονται στη στήλη Sig. Έτσι έχοντας ως επίπεδο σημαντικότητας $\alpha = 5\%$ προκύπτει ότι δεν απορρίπτουμε την υπόθεση ότι τα δεδομένα του χρόνου για κάθε έναν από τους δύο πληθυσμούς (αγόρια και κορίτσια) προέρχονται από πληθυσμούς που περιγράφονται ικανοποιητικά από την κανονική κατανομή (p-τιμή του Shapiro-Wilk=0.694 και 0.126 μεγαλύτερες του 0.05, αντίστοιχα για αγόρια και κορίτσια). Στο ίδιο συμπέρασμα καταλήγουμε χρησιμοποιώντας και τους γραφικούς τρόπους ελέγχους (Normal Q-Q Plot και Detrended Normal Q-Q Plot).

Tests of Normality

Φύλο		Kolmogorov-Smirnov(a)			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Χρόνος σε δευτερόλεπτα	Αγόρι	,116	19	,200(*)	,966	19	,694
	Κορίτσι	,144	16	,200(*)	,912	16	,126

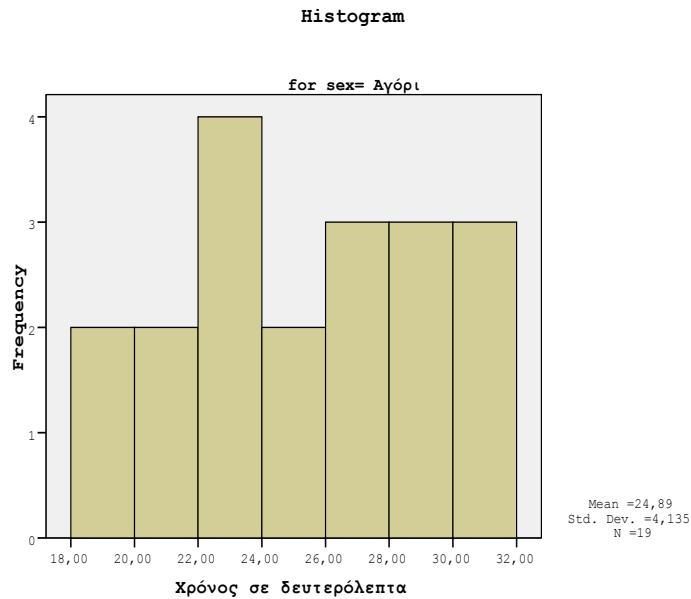
* This is a lower bound of the true significance. a Lilliefors Significance Correction

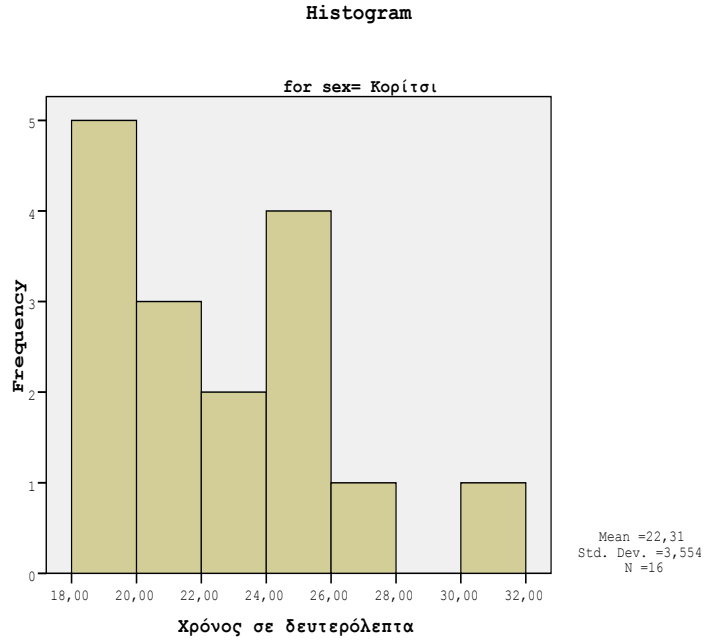
Στο πλαίσιο Test of Homogeneity of Variance δίνεται το στατιστικό τεστ του Levene για τον έλεγχο της υπόθεσης των ίσων διακυμάνσεων. Προκύπτει ότι δεν απορρίπτεται η υπόθεση των ίσων πληθυσμιακών διακυμάνσεων καθώς η p-τιμή του ελέγχου είναι ίση με $0.371 > 0.05$.

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Χρόνος σε δευτερόλεπτα	Based on Mean	,823	1	33	,371
	Based on Median	,823	1	33	,371
	Based on Median and with adjusted df	,823	1	32,896	,371
	Based on trimmed mean	,883	1	33	,354

Επιπλέον έχουμε το ιστόγραμμα και το φυλλογράφημα της μεταβλητής Χρόνος σε δευτερόλεπτα ως προς το φύλο.





Χρόνος σε δευτερόλεπτα Stem-and-Leaf Plot for sex= Αγόρι

Frequency	Stem & Leaf
2,00	1 . 88
7,00	2 . 0122334
7,00	2 . 5677889
3,00	3 . 002

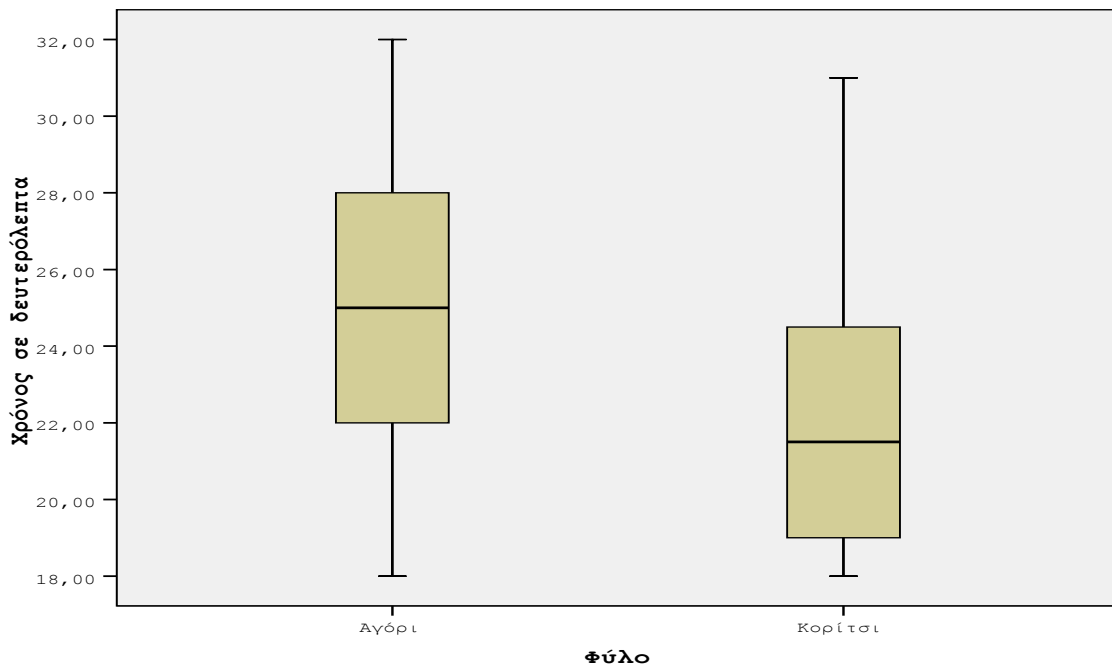
Stem width: 10,00
Each leaf: 1 case(s)

Χρόνος σε δευτερόλεπτα Stem-and-Leaf Plot for sex= Κορίτσι

Frequency	Stem & Leaf
5,00	1 . 89999
7,00	2 . 0112344
3,00	2 . 557
1,00	3 . 1

Stem width: 10,00
Each leaf: 1 case(s)

Στη συνέχεια παραθέτουμε το θηκόγραμμα (ως προς το φύλο) της μεταβλητής που περιγράφει το χρόνο που διανύουν τα παιδιά τα 100 μέτρα. Το θηκόγραμμα όπως παρατηρούμε μας δίνει τη δυνατότητα να συγκρίνουμε άμεσα τη διάμεσο, το 25^ο και 75^ο ποσοστιαίο σημείο, την μέγιστη και ελάχιστη παρατηρούμενη τιμή. Επιπλέον πιθανές ακραίες τιμές δηλώνονται με ένα ο, ενώ οι extreme (πολύ ακραίες) με ένα *. Στο συγκεκριμένο παράδειγμα προκύπτει ότι δεν έχουμε ακραίες τιμές, η διάμεσος, η μέγιστη και η ελάχιστη τιμή του χρόνου σε δευτερόλεπτα των αγοριών είναι μεγαλύτερη από τις αντίστοιχες τιμές για τα κορίτσια.



Επιπλέον δυνατότητες της διαδικασίας Descriptives

Ακολουθούμε την πορεία: Analyze → Descriptive Statistics → Descriptives, στο νέο παράθυρο διαλόγου που προκύπτει επιλέγοντας το πλαίσιο Save standardized values as variables καθίσταται δυνατή η αποθήκευση των τυποποιημένων τιμών (standardized

values) για τις μεταβλητές του καταλόγου Variable(s). Αν X_1, \dots, X_n είναι οι διαθέσιμες δειγματικές τιμές της μεταβλητής που δηλώθηκε στο πλαίσιο Variable τότε δημιουργείται μία νέα στήλη με τιμές Z_1, \dots, Z_n που υπολογίζονται από τη σχέση:

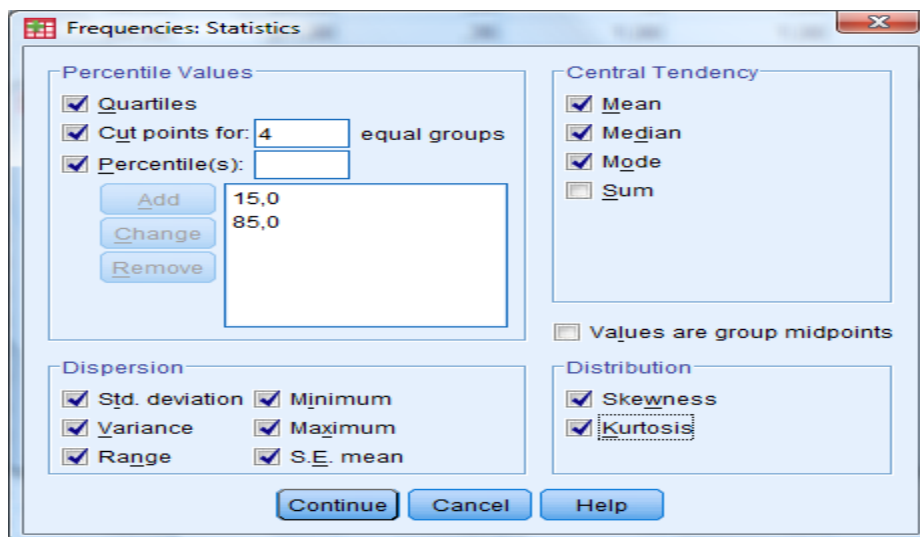
$$Z_i = \frac{X_i - \bar{X}}{S}$$

Οι τυποποιημένες τιμές ή Z-scores είναι μερικές φορές χρήσιμες για περαιτέρω ανάλυση. Με αυτές μπορούμε για παράδειγμα να συγκρίνουμε δείγματα από διαφορετικούς πληθυσμούς ή μετρήσεις μεταβλητών σε διαφορετικές μονάδες μέτρησης.

Επιπλέον δυνατότητες της διαδικασίας Frequencies

Αρχικά επιλέγουμε Analyze → Descriptive Statistics → Frequencies.

Από την επιλογή Statistics επιλέγοντας το πλαίσιο Cut points for: ζητούμε την εμφάνιση των σημείων εκείνων για το διαχωρισμό των δεδομένων σε τόσες ομάδες όσες ο αριθμός που θα δηλωθεί (π.χ. cut points for 4 equal groups), καθώς και την εμφάνιση π.χ. του 15^ο και 85^ο ποσοστιαίου σημείου. Τέλος επιλέγουμε το πλαίσιο Values are group midpoints αν οι τιμές των δεδομένων μας είναι το μέσο ενός διαστήματος. Για παράδειγμα αν πρόκειται για ηλικίες και για όσους είναι από 30-40 είχαμε καταχωρήσει στην αντίστοιχη μεταβλητή το μέσο του διαστήματος δηλαδή την τιμή 35, ενώ για εκείνους με ηλικία από 40-50 αντίστοιχα την τιμή 45 κ.ο.κ.



Επιπλέον, από την επιλογή Charts έχει νόημα μόνο η κατασκευή ιστογράμματος (histogram) ζητώντας παράλληλα να σχεδιαστεί και η κανονική καμπύλη (normal curve). Η επιλογή αυτή μας δίνει τη δυνατότητα να διαπιστώσουμε αν έχουμε ενδείξεις για αποκλίσεις από την κανονική κατανομή

